Computer Communications 56 (2015) 25–34

Contents lists available at ScienceDirect

Computer Communications

journal homepage: www.elsevier.com/locate/comcom

# Coarse-grained traffic matrix estimation for data center networks ☆

Zhiming Hu [a,*], Yan Qiao [a,b,*,1], Jun Luo [a]

[a] School of Computer Engineering, Nanyang Technological University, Singapore
[b] School of Information and Computer, Anhui Agricultural University, China

ARTICLE INFO

*Article history:*
Available online 28 February 2014

*Keywords:*
Data center networks
Network tomography
Traffic matrix estimation

ABSTRACT

We address the problem of estimating the real-time traffic flows in *data center networks* (DCNs), using the light-weight SNMP data. Unlike the problem of estimating the *traffic matrix* (TM) across origin–destination (OD) pairs in ISP networks, the traffic flows across servers or *Top of Rack* (ToR) switch pairs in DCNs are notoriously more irregular and volatile. Although numerous methods have been proposed in past several years to solve the TM estimation problem in ISP networks, none of them could be applied to DCNs directly. In this paper, we make the first step to solve the TM estimation problem in DCNs by leveraging the characteristics of prevailing data center architectures and decomposing the topologies of DCNs, which makes TM estimation problems in DCNs easy to handle. We also state a basic theory to obtain the aggregate traffic characteristics of these clusters unbiasedly. We propose two efficient TM estimation algorithms based on the decomposed topology and the aggregate traffic information, which improves the state-of-the-art tomography methods without requiring any additional instrumentation. Finally, we compare our proposal with a recent representative TM estimation algorithm through both real experiments and extensive simulations, the results show that, (i) the data center TM estimation problem could be well handled after the decomposition step, (ii) our two algorithms outperforms the former one in both speed and accuracy.



## 1. Introduction

As *data center networks* (DCNs) become increasingly central in cloud computing, both academic and industrial communities have started to explore how to better design and manage them. The main topics include network structure design [2–4], traffic engineering [5], capacity planning [6], anomaly detection [7], etc. However, until recently, very little is known about the characteristics of traffic flows within DCNs. For instance, how do traffic volumes exchanged by two servers or *Top of Rack* (ToR) switches vary with time? Which server communicates to other servers the most in DCNs? Actually the real-time *traffic matrix* (TM) across servers or ToR switches is a critical input to all above network designs and operations. Lack of this information hampers both research and practice.

With the increasing demands for the detailed flow level information of DCNs, a few works have studied the flow characteristics of the data centers in their hands [8–10]. However, the main barrier for them is the difficulty in flow data collection, for the flow level instrumentation is unavailable in most data centers. Besides, installing these additional modules requires substantial development and lots of administrative costs.

As the SNMP counters are ubiquitously available in all DCN devices, it is natural to ask if we could borrow from the well known tomography methods [11–13] and use link level information (such as SNMP bytes counters) to infer the TMs in DCNs. Unfortunately, both Kandula et al.'s experiments in real DCNs [10] and our testbed show that all existing tomography based methods (reviewed in Section 2) perform poorly in DCNs. This is due to the irregular flow behaviors and the large quantity of redundant paths between each pair of servers or ToR switches in these networks.

In this paper, we demonstrate that the prevailing DCN topologies (including conventional data center architecture [14], Fat-Tree [2], VL2 [3], etc.) can be divided into several clusters, and the complexity of original TM estimation problem can be reduced accordingly. Based on that, we design two efficient algorithms to infer, with high accuracy, the TMs (i) across these clusters and (ii) among ToR switches within each cluster. Then we verify their

☆ A preliminary version of the article has been published in IFIP Networking 2013 [1].

* Corresponding authors. Address: School of Computer Engineering, Nanyang Technological University, Singapore 639798, Singapore. Tel.: +65 97716942.
*E-mail addresses:* zhu007@ntu.edu.sg (Z. Hu), qiaoyan101@gmail.com (Y. Qiao), junluo@ntu.edu.sg (J. Luo).

[1] The work was done when she was a postdoc fellow of SCE NTU.

http://dx.doi.org/10.1016/j.comcom.2014.02.016
0140-3664/

performance in our experiments. More specifically, this paper makes the following contributions to the field of data center networking.

We decompose DCN topology into several clusters to deal with the large quantity of paths between *origin–destination* (OD) pairs. By doing this, the complexity of the intractable inference problem can be dramatically reduced, and tomography methods may hence be applied. We also state a basic theory that the total traffics exchanged among clusters and within each cluster can both be unbiasedly inferred from the link loads on switches. Such aggregate traffic characteristics are of great significance for the network administrators. For instance, clusters with much more intra traffic may have been well designed, as the intra traffic often costs lower network and computational resources. And the administrators should pay more attention to the clusters that communicate a lot with other clusters, whose traffic may cause relative high network delay.

We propose two efficient algorithms to infer the detailed inter and intra clusters' TMs. The first algorithm, which is more appropriate to infer the TMs without explicit structures, utilizes the aggregate traffic information to calculate a hypothesis flow volume on each path and then refine the assignments by a least square problem. The second one models the inference problem as a state-space network which incorporates both the spatial and temporal structure of TM, and updates the states of TM elements whenever a new observation arrives.

Finally, we design several experiments on testbed and extensive simulations in *ns*-3 to validate the performances of our two proposals. Through comparing with a recent representative TM estimation method, the experiment results show that our two algorithms outperform the former algorithms in both accuracy and speed, especially for large scale TMs.

The rest of the paper is organized as follows: we survey the related works in Section 2, and present the problem formulation in Section 3. In Section 4, we present DCN topology decomposition principles. We propose two efficient TM estimation algorithms in Section 5 and Section 6, respectively and evaluate them through both experiments and simulations in Section 7. Finally we give a discussion in Section 8 and conclude our work in Section 9.

## 2. Related work

As DCN has recently emerged as an intriguing topic, there are numerous studies working on approaches for traffic engineering [5], anomaly detection [7], provisioning and capacity planning [6], etc. However, almost no existing work has devoted to the traffic measurement approaches, although the estimation of traffic flows is a critical input to all above network designs and operations.

Previous studies [8,9] have exploited the traffic characteristics within DCNs. The former focuses on cloud data centers that host Web services as well as those running MapReduce [15], while the latter considers more generic DCNs such as enterprise and campus data centers. Both of them collected packet traces by attaching a dedicated packet sniffer on the switches in data centers. It is an impractical solution to turn on the packet sniffers all the time since it will consume a lot of switch resources. Therefore, Benson et al. in [9] only selected a handful of locations at random per data center and installed sniffers on them.

Kandula et al. [10] studied the nature of data center traffic on a single MapReduce data center. They firstly measure the traffic on data center servers, providing socket level logs. They also question whether TM can be inferred from link counters by tomography methods in DCNs as they perform in the ISP counterpart? If they do so, the barrier to understand the traffic characteristics of data centers will be lowered from the expensive instrumentation to analyzing the more easily available SNMP link counters. Unfortunately, they show with their evaluations that tomography performs poorly for data center traffic, due to the following reasons.

- Most existing topography based methods model the traffic flows at the granularity of volumes exchanged by OD pairs, assuming that there is only one path between an OD pair and the routing matrix will always be constant over time. However, this assumption may be violated in DCNs. There are a great number of redundant paths in DCNs to deal with the congestion, and choosing which route depends on the particular scheduling strategy within the network.
- To address the under-determined problem in network topography, some methods make additional assumptions such as gravity traffic model [11] and rank minimization [12], both of which perform poorly in DCNs, since servers in data centers do not have the same behaviors as terminals in ISP networks.
- Methods that exploit the spatio-temporal structure of traffic flows [12] often have high time and space complexity, for the elements in the TM were estimated simultaneously under the global constrains. When TM is in large scale, these inference algorithms incur high time and space complexities. Moreover, when new observations or requirements arrive, they need to start over again.

In this paper, we aim at designing an efficient tool to infer the traffic flows in DCNs with high accuracy by the ubiquitous SNMP data collected by switches. With the our new but powerful tool, the data center administrators could learn the real-time network traffic details at any moment they need.

## 3. Problem formulation background

We consider a typical DCN as shown in Fig. 1, consisting of ToR switches, aggregation switches and core switches connecting with Internet. We can poll the SNMP MIBs on the network switches for bytes/packets-in and bytes/packets-out at granularities ranging from 1 min to 30 min. The SNMP data can also be interpreted as switch loads equals to the summation of volumes of flows that traverse the corresponding switches. While a *fine-grained* DCN TM indicates the traffic volumes exchanged between ToRs, we decompose a DCN into clusters and aim at only inferring *coarse-grained* TMs (among clusters and ToRs in each cluster) from switch loads.

We represent switches in the network as $S = \{s_1, s_2, \ldots, s_m\}$, where $m$ is the number of switches. Let $\mathbf{y} = \{y_1, y_2, \ldots, y_m\}$ denote the traffic loads collected by SNMP counters on the switches, and $\mathbf{x} = \{x_1, x_2, \ldots, x_n\}$ denote the traffic flow volumes on the paths between ToR switch pairs, where $n$ is the number of all available paths in DCN. $x_i(t)$ and $y_j(t)$ represent the corresponding traffic at discrete time $t$. The correlation between $\mathbf{x}(t)$ and $\mathbf{y}(t)$ can be formulated as

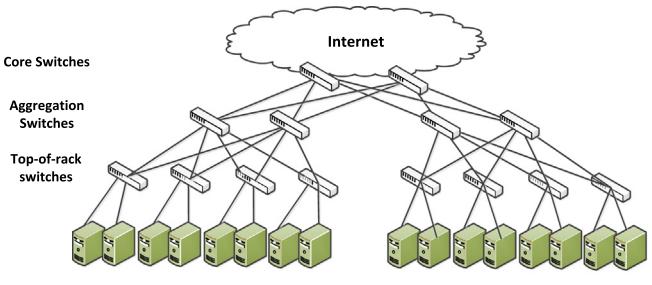


**Fig. 1.** An example of conventional DCN architecture (adopted from Cisco [14]).